SCIENCE CHINA
Life Sciences

THEMATIC ISSUE: Computational life sciences
• RESEARCH PAPER •

November 2014 Vol.57 No.11: 1054–1063
doi: 10.1007/s11427-014-4745-8

Disease gene identification by using graph kernels and Markov random fields

CHEN BoLin[1], LI Min[2], WANG JianXin[2] & WU FangXiang[1,3*]

[1]*Division of Biomedical Engineering, University of Saskatchewan, Saskatoon S7N 5A9, Canada;*
[2]*School of Information Science and Engineering, Central South University, Changsha 410083, China;*
[3]*Department of Mechanical Engineering, University of Saskatchewan, Saskatoon S7N 5A9, Canada*

Received May 21, 2014; accepted July 14, 2014; published online October 14, 2014

Genes associated with similar diseases are often functionally related. This principle is largely supported by many biological data sources, such as disease phenotype similarities, protein complexes, protein-protein interactions, pathways and gene expression profiles. Integrating multiple types of biological data is an effective method to identify disease genes for many genetic diseases. To capture the gene-disease associations based on biological networks, a kernel-based Markov random field (MRF) method is proposed by combining graph kernels and the MRF method. In the proposed method, three kinds of kernels are employed to describe the overall relationships of vertices in five biological networks, respectively, and a novel weighted MRF method is developed to integrate those data. In addition, an improved Gibbs sampling procedure and a novel parameter estimation method are proposed to generate predictions from the kernel-based MRF method. Numerical experiments are carried out by integrating known gene-disease associations, protein complexes, protein-protein interactions, pathways and gene expression profiles. The proposed kernel-based MRF method is evaluated by the leave-one-out cross validation paradigm, achieving an AUC score of 0.771 when integrating all those biological data in our experiments, which indicates that our proposed method is very promising compared with many existing methods.

disease gene identification, data integration, Markov random field, graph kernel, Bayesian analysis

Citation: Chen BL, Li M, Wang JX, Wu FX. Disease gene identification by using graph kernels and Markov random fields. Sci China Life Sci, 2014, 57: 1054−1063, doi: 10.1007/s11427-014-4745-8

The availability of large-scale biological networks provides an opportunity to comprehensively identify disease genes of many genetic diseases, by synergizing evidences from multiple types of data sources. Various algorithms [1–6] have been developed to identify human disease genes based on the strategy of multiple data integration.

However, challenges still exist due to the following reasons. Firstly, there are many levels of controls along paths from genotypes to phenotypes [7], resulting in the indirect relationship between genotypes and phenotypes [8]. Secondly, different biological data are heterogeneous and describe relationships of molecular entities in various levels. It is not a trivial task to design a good algorithm that combines those data appropriately. Thirdly, many data integration methods simply assume that disease genes of similar diseases exhibit dense clusters in the integrated networks, but ignore the fact that those networks are built independently from the description of gene-disease association relationships.

The Markov random field (MRF) model proposed by Deng et al. [9,10] for predicting yeast protein functions provides a good framework to integrate multiple biological networks. The issue of protein function prediction is formulated as a Bayesian labeling problem, where the function

*Corresponding author (email: faw341@mail.usask.ca)

labels follow a Gibbs distribution. A binary logistic regression is employed to estimate parameters from known observations, and a Gibbs sampling approach is developed to generate final predictions. Advantages of the MRF method include its simplicity, its ability to explore contributions of direct neighbors, and its flexibility to integrate multiple types of datasets.

Although the issue of yeast protein function prediction is similar to the issue of human disease gene identification, the method of Deng et al. [9,10] cannot be directly employed to identify human disease genes. Parameters of the MRF model cannot be estimated precisely due to the limited observations of human disease genes, which make predictions of their method unreliable. Kourmpetis et al. [11] then proposed a Bayesian MRF method to estimate parameters iteratively together with the update of posterior probabilities of function labels during the Gibbs sampling process. However, their method uses another predefined scaling parameter γ, a Z matrix and a multivariate normal distribution to perform the estimation, which makes the method a little complex. Ma et al. [5] proposed a combining gene expression and protein interaction data (CGI) method to identify genes responsible for similar phenotypes or traits, motivated from the MRF model of Deng et al. [9,10]. Similarity metric defined by the diffusion kernel is also compared with those by direct neighbors and shortest paths, where predictions from the diffusion kernel are greatly improved. However, the CGI method mainly uses gene expression profiles to group genes with similar characters. Protein interaction data are only employed to calibrate predictions. It is not clear how to integrate other types of biological data by using their method. Lee et al. [12] developed a kernel logistic regression (KLR) method for predicting yeast protein functions by combining advantages of both the MRF model and diffusion kernels. Although its predictive accuracy is higher than that of the original MRF method of Deng et al. [9,10], the parameter estimation problem still exists if the KLR method is employed to identify human disease genes. Other forms of MRF methods can be found in [13–15].

Graph kernels, on the other hand, have shown their powers for interpreting complex relationships of vertices in biological networks [16–18]. A kernel-based algorithm often yields better performance than those using direct neighbors or shortest paths under the same condition. In papers [19,20], we have developed a modified MRF model for human disease gene prioritization. In this study, we further propose a kernel-based MRF algorithm for identifying disease genes from multiple types of data by combining the MRF model and graph kernels. The kernel-based MRF algorithm is different from the methods proposed in [19,20] in the following four aspects. Firstly, a novel weighted MRF method is developed for incorporating different graph kernels. Secondly, a new parameter estimation method is designed for the kernel-based MRF method based on global characters of biological networks. Thirdly, an improved Gibbs sampling strategy is proposed which takes the weight value of neighbors into consideration, rather than simply counting the number of neighbors attributed specific values. Finally, the kernel-based MRF method is extended to integrate multiple types of data sources, such as protein-protein interaction (PPI) networks, pathway co-existing networks and gene co-expression networks. We show that the kernel-based MRF algorithm can significantly improve the accuracy of disease gene identification compared with existing methods.

1 Methods

1.1 Problem statement

Suppose a human genome consists of a set of N genes $\{g_1, g_2, ..., g_N\}$. Some of them have already been known to be associated with r genetic diseases, while associations of most others are still not known and need to be determined. Let $\{D_1, D_2, ..., D_r\}$ be those r associated diseases. Each D_i consists of a set of known disease genes of the ith disease. Hence, the number of all those known disease genes equals $m = |D_1 \cup D_2 \cup \cdots \cup D_r|$, where $|*|$ is the cardinality of the set. Without loss of generality, let $g_{n+1}, g_{n+2}, \ldots, g_{n+m}$ be those known disease genes, and $g_1, g_2, \ldots, g_n$ be all others, where $N=n+m$.

For a specific disease, let $x = (x_1, x_2, \ldots, x_{n+m})$ be a vector of binary variables (i.e., taking values zero or one) defined on all genes, where $x_i = 1$ represents gene g_i to be a disease gene of the disease and $x_i = 0$ otherwise. The purpose of disease gene identification is to predict values of $x^{\text{miss}} = (x_1, x_2, \ldots, x_n)$ from current known values $x^{\text{obs}} = (x_{n+1}, x_{n+2}, \ldots, x_{n+m})$. To achieve this, a vector of random variables $X = (X_1, X_2, \ldots, X_N)$ is defined corresponding to x, where $P(X_i = x_i)$ represents the probability that $X_i = x_i$. The objective is to find the posterior probability of $X_1, X_2, \ldots, X_n$ conditional on known disease genes,

$$P(X_1, X_2, \ldots, X_n | X_{n+1}, \ldots, X_{n+m}). \quad (1)$$

1.2 Markov random field

Let $G = (V, E)$ be a graph with N vertices and $X = (X_1, X_2, \ldots, X_N)$ be a vector of random variables defined on V. The vector X is said to be a MRF on G if and only if the following two conditions are satisfied:

(i) Positivity: $P(X_i) > 0$, $\forall X_i \in \mathcal{X}$,

(ii) Markovianity: $P(X_i|X_{[-i]}) = P(X_i|X_{N(i)})$,

where $\mathcal{X}$ are the set of all possible outcomes of X_i, $X_{[-i]}$ is the collection of random variable $(X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_N)$, and $X_{N(i)}$ is the collection of all X_j for $j \in N(i)$, where $N(i)$ is the set of all neighbors of vertex *i* in *G*. The neighborhood structure of graph *G* is denoted as $\mathcal{N}$. The Markovianity indicates that the probability of X_i is conditionally independent on all other X_k except the value of its neighbors. A joint event $\{X_1 = x_1,\ldots,X_N = x_N\}$, abbreviated as $X = x$, is a realization of *X*, where $x = (x_1, x_2,\ldots, x_N)$ is called a configuration of *X*.

One of the key features that facilitate the practical usage of MRF is its equivalence with the Gibbs random filed, which is established by the Hammersley-Clifford theorem [21,22]. According to the theorem, *X* is a MRF on *V* w.r.t. $\mathcal{N}$ if and only if the probability distribution of $P(X)$ follows a Gibbs distribution. The Gibbs distribution has a form of

$$P(X = x) = Z^{-1}e^{-U(x)/T}, \tag{2}$$

where $Z = \sum_{x\in\mathcal{X}} e^{-U(x)/T}$ is a normalizing constant called partition function, *T* is a global control constant called temperature, which is often assumed to be 1 unless otherwise stated, and $U(x)$ is called the energy function, which can be decomposed as a sum over all cliques in *G* [23], in the form

$$\begin{aligned} U(x) &= \sum_{c\in C} V_c(x) \\ &= \sum_{\{i\}\in C_1} V_1(x_i) + \sum_{\{(i,j)\}\in C_2} V_2(x_i, x_j) + R_n(x), \end{aligned} \tag{3}$$

where $V_i(x)$ is the clique potential of C_i (the set of *i*th order cliques in *G*), $R_n(x)$ represents those higher order terms. A special case of MRF is the Ising model that only considers up to the second order of cliques [24], which is also the same as many existing MRF methods did in [10,11,20].

The practical valuation of the Hammersley-Clifford theorem is that it gives a simple way to specify the probability $P(X)$ by using those clique potentials. Suppose X_i is a binary random variable (i.e., taking values zero or one). Let $V_1(x_i) = -\alpha\cdot x_i$, $V_2(1,1) = -\beta_{11}$, $V_2(1,0) = V_2(0,1) = -\beta_{10}$, and $V_2(0,0) = -\beta_{00}$. Let N_{11}, N_{10} and N_{00} be the number of edges whose two endpoints have both the attribute values of 1, one attribute value of 1 and the other value of 0, and both the attribute values of 0, respectively. Then the energy function (3) of the MRF can be written as

$$\begin{aligned} U(x) &= -\alpha\sum_{i\in V} x_i - \beta_{11}N_{11} - \beta_{00}N_{00} - \beta_{10}N_{10} \\ &= -\alpha\sum_{i\in V} x_i - \beta_{11}\sum_{\{(i,j)\}\in E} x_i x_j - \beta_{00}\sum_{\{(i,j)\}\in E}(1-x_i)(1-x_j) \\ &\quad -\beta_{10}\sum_{\{(i,j)\}\in E}[x_i(1-x_j) + (1-x_i)x_j], \end{aligned} \tag{4}$$

where $\theta = (\alpha, \beta_{11}, \beta_{10}, \beta_{00})$ are parameters.

To generate predictions from a MRF model, the value of parameter θ is necessary, which is generally unknown. The most natural approach to estimate θ is through the maximum likelihood estimation (MLE) method. However, the MLE method is often intractable in this situation, since the normalizing partition function *Z* is also a function of parameters. Fortunately, the pseudo-likelihood approach and the Gibbs sampling process provide a solution for estimating parameters and generating predictions from a MRF model.

Firstly, for estimating parameters, suppose parameters $\theta = (\alpha, \beta_{11}, \beta_{10}, \beta_{00})$ of (4) are given. Then fixing the value of X_i, the energy function of (4) can be rewritten as

$$\begin{aligned} &U(X_i = 1, X_{[-i]}|\theta) \\ &\quad = U(X_{[-i]}|\theta) - \alpha - \beta_{11}\sum_{j\in N(i)} x_j - \beta_{10}\sum_{j\in N(i)}(1-x_j) \end{aligned} \tag{5}$$

and

$$\begin{aligned} &U(X_i = 0, X_{[-i]}|\theta) \\ &\quad = U(X_{[-i]}|\theta) - \beta_{10}\sum_{j\in N(i)} x_j - \beta_{00}\sum_{j\in N(i)}(1-x_j) \end{aligned} \tag{6}$$

respectively, where $U(X_{[-i]}|\theta)$ represents the energy contributed by all cliques that do not contain vertex *i*.

Hence, according to the Bayes' theorem [25] and (2), (5) and (6), we have

$$\begin{aligned} P(X_i = 1|X_{[-i]},\theta) &= \frac{P(X_i = 1, X_{[-i]}|\theta)}{P(X_i = 1, X_{[-i]}|\theta) + P(X_i = 0, X_{[-i]}|\theta)} \\ &= \frac{e^{-U(X_i=1, X_{[-i]}|\theta)}}{e^{-U(X_i=1, X_{[-i]}|\theta)} + e^{-U(X_i=0, X_{[-i]}|\theta)}}. \end{aligned} \tag{7}$$

The log-odds of the probability $P(X_i = 1|X_{[-i]},\theta)$ is

$$\log\frac{P(X_i = 1|X_{[-i]},\theta)}{1 - P(X_i = 1|X_{[-i]},\theta)} = \alpha + \beta_1 M_{i1} + \beta_0 M_{i0}, \tag{8}$$

where $\beta_1 = (\beta_{11} - \beta_{10})$, $\beta_0 = (\beta_{10} - \beta_{00})$, and M_{i1}, M_{i0} are the number of neighbors of gene *i* whose x_j are attributed with value of 1 and 0 in *G*, respectively. Those parameters of the MRF method can be estimated by using the standard MATLAB function *glmfit*().

Secondly, for the Gibbs sampling process, it is a type of

Markov Chain Monte Carlo (MCMC) algorithm. Given a set of probabilities $X^{(t)}$ at time *t*, it iteratively updates the value of *X* according to the univariate conditional distribution $P(X_i = 1 | X_{[-i]}, \theta)$ as follows:

$$\begin{aligned} X_1^{(t+1)} &\Leftarrow P(X_1 | X_2^{(t)}, \ldots, X_n^{(t)}, X^{\text{obs}}, \theta) \\ X_2^{(t+1)} &\Leftarrow P(X_2 | X_1^{(t+1)}, X_3^{(t)}, \ldots, X_n^{(t)}, X^{\text{obs}}, \theta) \\ X_3^{(t+1)} &\Leftarrow P(X_3 | X_1^{(t+1)}, X_2^{(t+1)}, X_4^{(t)}, \ldots, X_n^{(t)}, X^{\text{obs}}, \theta) \\ &\vdots \\ X_n^{(t+1)} &\Leftarrow P(X_n | X_1^{(t+1)}, \ldots, X_{n-1}^{(t+1)}, X^{\text{obs}}, \theta) \end{aligned} \quad (9)$$

where $X^{\text{obs}} = (X_{n+1}, \ldots, X_{n+m})$. The Gibbs sampling process always uses the most recent values of X_i to update successive variables. The sequence $X^{(1)}, X^{(2)}, X^{(3)}, \ldots$ clearly forms a Markov chain.

1.3 Graph kernels

Kernels provide a general framework to represent data in the form of pairwise similarities. Generally, two mathematical conditions need to be satisfied that make a function *k* serving as a kernel: (i) it must be symmetric $(k(x_i, x_j) = k(x_j, x_i))$ and (ii) positive semi-definite. Mathematically, for any kernel function *k* on a space $\mathcal{X}$, there exists a Hilbert space $\mathcal{H}$ and a mapping $\phi: \mathcal{X} \to \mathcal{H}$, such that

$$k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle, \text{ for any } x_i, x_j \in \mathcal{X}, \quad (10)$$

where $\langle u, v \rangle$ represents the dot product in the Hilbert space between any two points $u, v \in \mathcal{H}$.

The definition of a kernel is a critical component of any kernel method, since it defines how an algorithm "sees" the data. The graph representation of a biological network is often used to describe local topological relationships, which is often not enough to capture the distant relationships among biomolecules. Alternately, a graph kernel based representation provides a solution for this by considering global topological structures [16,18].

One of the most commonly used graph kernel of G is the Laplacian exponential diffusion (LED) kernel [16], where the kernel matrix is defined as

$$\begin{aligned} K_{\text{LED}} = e^{-\lambda L} &= \lim_{m \to \infty} \left(I - \frac{\lambda L}{m} \right)^m \\ &= I - \lambda L + \frac{(\lambda L)^2}{2!} - \frac{(\lambda L)^3}{3!} + \cdots \end{aligned} \quad (11)$$

where $L = D - A$ is the Laplacian matrix of the graph *G*, *A* is the adjacency matrix, and *D* is a diagonal matrix with the *i*th diagonal element $d(i)$ being the degree of the vertex *i* and all off-diagonal elements being 0. The parameter λ controls the magnitude of the diffusion, which is often chosen as a very small number. In this study, we take $\lambda = 0.04$ as [7] suggested.

A diffusion kernel defines the similarity of biomolecule pairs by considering all pairwise relationships within a network. However, diffusion kernels between different biological networks may not be comparable when a method needs to integrate multiple data sources. To overcome this problem, Chen et al. [7] propose a measure, called DKPC, to normalize pairwise similarities based on their relative strengths among all similarities within a network. The DKPC value between a vertex pair *i* and *j* is defined as

$$DKPC(i, j) = \frac{\left| \{(s,t) | K_{st} \geqslant K_{ij}\} \right|}{\left| \{(s,t) | K_{st} \geqslant 0\} \right|}, \quad (12)$$

where K_{ij} is a similarity value of vertex pair *i* and *j* in a kernel matrix. A smaller value of $DKPC(i, j)$ indicates that two vertices *i* and *j* are more similar. However, in the K_{LED} matrix, it uses larger values to represent relationships of vertices are more similar. To be consistent, we use the complementary value $\overline{DKPC}_{ij} = 1 - DKPC(i, j)$ to represent the normalized similarity between vertex pair *i* and *j* that is obtained from the DKPC method hereafter.

Generally, the above two kernels are strongly related to the degree of individual vertices, where the kernel value between two high degree vertices is significantly different from that between two low degree vertices. To make the strength of individual vertices comparable, we propose a Markov exponential diffusion (MED) kernel in this study by replacing the Laplacian matrix *L* in (11) with a Markov matrix *M*, which consists of nonnegative real numbers with each row and column summing to 1. The MED kernel matrix is defined as follows:

$$\begin{aligned} K_{\text{MED}} = e^{-\lambda M} &= \lim_{m \to \infty} \left(I - \frac{\lambda M}{m} \right)^m \\ &= I - \lambda M + \frac{(\lambda M)^2}{2!} - \frac{(\lambda M)^3}{3!} + \cdots \end{aligned} \quad (13)$$

where $M = (N \cdot I - D + A)/N$ and *N* is the number of vertices in the network.

1.4 Kernel-based MRF method

Let $K_{N \times N}$ be a kernel matrix derived from a biological network, where all diagonal elements are set to be zero (since the purpose of disease gene identification is to obtain a set of novel candidate genes according to the knowledge of known disease genes, the similarity metrics between a

gene and itself are neglected). Let $p=(p_1,p_2,\ldots,p_N)$ be a vector of probabilities, where p_i represents the probability of $X_i=1$ conditional on all other variables $X_{[-i]}$ given the parameter θ. We propose the kernel-based MRF method in three steps as follows.

Firstly, let $x=(x_1,x_2,\ldots,x_N)$ be a set of configuration obtained from *p*. In the KLR method of Lee et al. [12], the weighted number of neighbors whose x_j values are attributed with 1 and 0 for gene *i* are defined as

$$M_{i1}^{w}=\sum_{j=1}^{N}K_{ij}\cdot x_j \quad \text{and} \quad M_{i0}^{w}=\sum_{j=1}^{N}K_{ij}\cdot(1-x_j) \qquad (14)$$

respectively, where K_{ij} is the entry in the *i*th row and *j*th column of the matrix $K_{N\times N}$. It should be noticed that the values of M_{i1}^{w} and M_{i0}^{w} are highly dependent on values of all x_j, which are randomly generated from those p_j. To reduce the dependence, the x_j in (14) can be replaced directly by using those p_j. Thus, the improved weighted number of neighbors can be written as

$$M_{i1}^{w'}=\sum_{j=1}^{N}K_{ij}\cdot p_j \quad \text{and} \quad M_{i0}^{w'}=\sum_{j=1}^{N}K_{ij}\cdot(1-p_j). \qquad (15)$$

The log-odds of the probability $P(X_i=1|X_{[-i]},\theta)$ in weighted form is then defined as

$$\log\frac{P(X_i=1|X_{[-i]},\theta)}{1-P(X_i=1|X_{[-i]},\theta)}=\alpha+\beta_1 M_{i1}^{w'}+\beta_0 M_{i0}^{w'}. \qquad (16)$$

Secondly, an improved Gibbs sampling method is proposed that can iteratively estimate and update parameters θ simultaneously with the change of posterior probabilities. Suppose a prior probability of $p^{(0)}$ is given for all vertices. A set of prior configuration $x^{(0)}=(x_1^0,\ldots,x_N^0)$ can be randomly generated according to the prior probability. Then the pseudo-likelihood parameter estimation method can be performed on the whole network, including those known vertices and those unknown vertices, by using (16). Once those parameters are obtained in this iteration, the posterior probabilities of each vertex p_i can then be updated according to (16) as well. Repeat this process for many times until both of them are stable. The step-by-step description of this Gibbs sampling procedure is given as follows.

S1. Initialization:

Let $t=0$. Initialize the prior probabilities for unknown vertices $(p_1^{(0)},p_2^{(0)},\ldots,p_n^{(0)})$ and known vertices $p^{\text{obs}}=(p_{n+1},p_{n+2},\ldots,p_{n+m})$ respectively.

S2. Parameter estimation:

Assign a configuration of $x^{(t)}=(x_1^{(t)},\ldots,x_N^{(t)})$ and calculate the values of $M_{i1}^{w'}$ and $M_{i0}^{w'}$ according to the value of $p^{(t)}=(p_1^{(t)},\ldots,p_n^{(t)},p^{\text{obs}})$. Estimate parameters $\theta^{(t)}$ based on (16).

S3. Gibbs sampling:

$$p_1^{(t+1)}\Leftarrow P(X_1=1|p_2^{(t)},\ldots,p_n^{(t)},p^{\text{obs}},\theta^{(t)})$$
$$p_2^{(t+1)}\Leftarrow P(X_2=1|p_1^{(t+1)},p_3^{(t)},\ldots,p_n^{(t)},p^{\text{obs}},\theta^{(t)})$$
$$\vdots$$
$$p_n^{(t+1)}\Leftarrow P(X_n=1|p_1^{(t+1)},\ldots,p_{n-1}^{(t+1)},p^{\text{obs}},\theta^{(t)})$$

S4. Let $t=t+1$, and go to S2, until t is larger than a predefined iteration step.

The details of how this predefined iteration step is set are given in the Experimental design section. The improved Gibbs sampling procedure above is different from existing MRF methods [10,13,23] in two aspects. First, parameters of the improved method are estimated according to the configuration and the posterior probability of the whole network, while most existing MRF methods are based on subnetworks that consist of only known vertices. Ignoring the majority of unknown vertices makes the value of M_{i1} and M_{i0} (or $M_{i1}^{w'}$ and $M_{i0}^{w'}$ in this study) inaccurate, and then parameters cannot be estimated precisely. Predictions become unreliable if those inaccurate parameters are used to identify human disease genes. Second, many existing MRF methods estimate parameters only once. Parameters are then fixed during the entire Gibbs sampling process [10,13,23]. This is very dangerous if parameters are not estimated precisely. In our method, parameters are updated iteratively together with the change of all posterior probabilities. The Gibbs sampling process always takes the most updated parameters to estimate posterior probabilities for all unknown vertices, which is expected to generate more reliable predictions.

Finally, the proposed MRF method is extended for incorporating multiple types of biological networks. Suppose there are L networks $H=(H^1,\ldots,H^L)$, where vertices represent genes and edges represent specific biological relationship between vertices. Eq. (16) can be easily extended as

$$\log\frac{P(X_i=1|X_{[-i]},\theta)}{1-P(X_i=1|X_{[-i]},\theta)}=\alpha+\sum_{l=1}^{L}\left[\beta_1^l M_{i1}^{w'l}+\beta_0^l M_{i0}^{w'l}\right] \qquad (17)$$

by simply summing the effect of the weighted number of neighbors $M_{i1}^{w'l}$ and $M_{i0}^{w'l}$ for gene *i* from all *L* networks, where $\theta=(\alpha,\beta_1^1,\beta_0^1,\ldots,\beta_1^L,\beta_0^L)$ are parameters. The contribution of a network H^l can be adjusted through the value of β_1^l and β_0^l accordingly. The improved Gibbs sampling procedure can be easily performed by replacing

(16) with (17), when estimating parameters and updating posterior probabilities during the iterations.

1.5 Experimental design

1.5.1 Data Sources

Known gene-disease associations are collected from the Morbid Map list of the Online Mendelian Inheritance in Man (OMIM) [26]. Goh et al. [27] classify all those diseases into 22 primary disease classes, including a 'multiple' class and an 'unclassified' class. The dataset consists of 1284 diseases and 1777 disease genes. In this study, we choose those disease classes that consist of at least 30 genes and exclude the 'multiple' class, the 'unclassified' class, the 'cancer' class and the 'neurological' class due to the lack of their class evidence and the class heterogeneity [27]. The final dataset consists of 815 genes in 12 disease classes.

Two sets of protein complexes are collected from the database of CORUM [28] and PCDq [29], which contain 1677 and 1103 protein complexes that consist of at least two proteins, respectively. All those protein complexes are employed to assign the prior probabilities for unknown vertices.

Three PPI networks are derived from the database of HPRD (Release 9) [30], BioGrid (Release 3.2.108) [31] and IntAct (downloaded on Jan 26, 2014) [32], respectively. Duplicated edges and loop edges are deleted. The HPRD PPI network consists of 9465 vertices and 37039 edges. The BioGrid PPI network consists of 15298 vertices and 127612 edges. The IntAct PPI network consists of 13449 vertices and 63825 edges. These PPI networks have been widely used to identify protein complexes [33–36] or essential proteins [37–39] and thus can be considered to be reliable data.

Pathway datasets are obtained from the database of KEGG [40], Reactome [41], PharmGKB [42] and PIN [43], which contain 280, 1469, 99 and 2679 pathways, respectively. A pathway co-existing network is constructed by taking individual proteins/genes as vertices. Edges are constructed between two vertices if they co-exist in any pathway.

A gene co-expression network is constructed by using the dataset of BioGPS (GSE1133) [44,45]. It contains 79 human tissues in duplicates, which are measured by using the Affymetrix U133A array. Pairwise Pearson correlation coefficients (PCC) are calculated and a pair of genes are linked by an edge if the PCC value is larger than 0.5, similar to the method used in [7,27].

Overall, five biological networks are constructed and all protein (or gene) IDs are mapped onto the form of the gene symbol. In order to test the performance of multiple data integration of our method, we select those genes that appear at least in four networks. The final datasets consist of 7311 human genes, 815 out of which are known to be associated with 12 disease classes.

1.5.2 Estimating a prior probability

To perform a Gibbs sampling procedure, a set of prior probabilities for all vertices is needed. Generally, the values of those prior probabilities do not have significant effect on the final stabled state of a Markov chain if enough iterations are performed. However, a good prior does help to reduce the time of iterations to achieve the stable state.

For those known disease genes, the prior probability of $p^{\text{obs}} = (p_{n+1}, \ldots, p_{n+m})$ can be assigned determinedly according to known gene-disease associations. The value of p_j, $n+1 \leqslant j \leqslant n+m$ equals 1 or 0 depending on the analyzed disease class and those known gene-disease associations.

For those unknown disease genes, since genes that encode proteins in a same complex tend to associate with similar diseases, we estimate their prior probabilities according to the protein complex information, similarly to the method used in Deng et al. [9,10]. For a gene g_i that encodes protein in a complex, let

$$\hat{p}_i = \frac{A}{B} \tag{18}$$

be the prior probability, where A is the number of disease genes for a specific disease in the complex, and B is the number of all disease genes in the complex. If a gene appears in multiple protein complexes, we use the maximum value as the prior probability for the gene. For those genes that do not belong to any protein complex, let

$$\hat{p}_i = \frac{C}{D} \tag{19}$$

as the prior probability, where C is the number of all currently known disease genes for the specific disease, and D is the total number of genes in human genome.

1.5.3 Specifying an iteration loop

During the Gibbs sampling procedure, a "burn-in period" and a "lag period" often need to be specified. The "burn-in period" is the period that a Markov chain takes to become stabilized. Simulation results in this period are discarded to reduce the effect of initial prior probabilities. The "lag period" is the period that needs to reduce the dependence of the Markov process. The posterior probabilities in this period are estimated by averaging simulation results during individual lag steps. In paper [20], we have shown that an additional "prediction period" is helpful to generate more stable and reliable predictions, which is the period used to generate final prediction by averaging all simulation results during this period.

In this study, the "burn-in period" takes 100 steps, the "lag period" takes 90 steps and the "prediction period" takes

100 steps. Simulation results are averaged every 10 steps in the "lag period". There are 1100 steps in total for simulations.

1.5.4 Validation method and evaluation criteria

The leave-one-out cross validation paradigm is employed to evaluate the proposed method. For each known disease gene with at least one annotated interaction partner in a biological network, we assume it is an unknown gene and predict its posterior probability by the proposed method. The receiver operating characteristic (ROC) curve is employed as one of the evaluation criteria, which shows the relationship between the true positive rate (TPR) and the false positive rate (FPR) by varying the threshold for declaring positives. The area under the ROC curve (AUC) is also employed to show an overall performance of an algorithm. The negative control set consists of known disease genes that do not belong to the current disease class, and they are also validated by using the leave-one-out cross validation paradigm. The application of the AUC score as the evaluation criterion is due to the fact that it is widely accepted by most researchers.

We compare the proposed method with four existing algorithms: (i) the random walk with restart (RWR) algorithm proposed by Köhler et al. [46]; (ii) the data integration rank (DIR) algorithm proposed by Chen et al. [7]; (iii) the original MRF method proposed by Deng et al. [10] (denoted as MRF-Deng hereafter) and (iv) our previous improved MRF method for identifying human disease genes [20] (denoted as IMRF hereafter). The RWR algorithm [46] is a typical data integration method that uses a mixed network, where vertices and edges of several biological networks are simply merged together, while our proposed method integrates those networks separately. The comparison between those two algorithms can show which manner of multiple data integration is better. The DIR algorithm [7] has a very good performance in terms of multiple data integration. It also employs diffusion kernels to integrate different networks separately, which yields better performance than many other data integration methods [7]. The comparison with the other two existing MRF methods demonstrates how much improvement can be obtained from the proposed method as well.

1.5.5 Decision score and declaration of positives

Different disease classes consist of different numbers of known disease genes, and thus the prediction results may not be good if a global threshold is used for all classes. Although one can directly use the posterior probabilities obtained from the Gibbs sampling to select candidate disease genes, we suggest using a percentage as a decision score to generate the finial predictions. Let $p^{(T)}=(p_1^{(T)}, p_2^{(T)},\ldots, p_n^{(T)})$ be the set of final posterior probabilities for a specific disease class. The decision score q_i of vertex *i* is defined as

$$q_i = \frac{\left|\{s \,|\, p_i^{(T)} \geqslant p_s^{(T)}\}\right|}{n}.$$

The greater the decision score is for a gene, the more likely it is to associate with specific disease. All the ROC curves and the AUC scores of the proposed method are calculated according to the decision score hereafter.

2 Results

2.1 Stability and reliability of the kernel-based MRF method

We first investigate the stability and reliability of the kernel-based MRF method, by comparing the Markov processes of the proposed method and the MRF-Deng method. Figure 1 illustrates the variation of posterior probabilities over iteration steps and the final posterior probability distribution for the above two methods.

Firstly, by comparing Figure 1A and C, we can clearly find that the kernel-based MRF method is more stable than the MRF-Deng method. The change of posterior probability of the front method converges quickly and stays at a stable state.

Here, the variation of posterior probabilities for two consecutive steps is calculated from

$$Q(t) = \sum_{i=1}^{n} \left(p_i^{(t)} - p_i^{(t-1)} \right)^2, \tag{20}$$

where $p_i^{(t)}$ is the posterior probability $P(X_i = 1 \,|\, X_{[-i]}, \theta)$ of g_i obtained in the *t*th iteration.

Secondly, predictions of the kernel-based MRF method are more reasonable compared with the MRF-Deng method. The parameters of the MRF-Deng method are estimated from subnetworks of known vertices, which may only be feasible when the subnetwork is large enough for estimating parameters precisely. When the MRF-Deng method is employed directly to identify human disease genes, there are approximately 25.82% unknown genes that are predicted as disease genes with a probability large than 0.95. This is unreasonably high in practice, since it contains too many false positive predictions. Figure 1D shows the final posterior probability distribution of the MRF-Deng method as an example.

On the other hand, the kernel-based MRF method works very well. Taking the endocrine disease class for example, which is illustrated in Figure 1B, most genes are predicted with a probability small than 0.1. Only a few significant vertices are predicted with higher probabilities. Predictions of the kernel-based MRF method are more reliable than the MRF-Deng method.

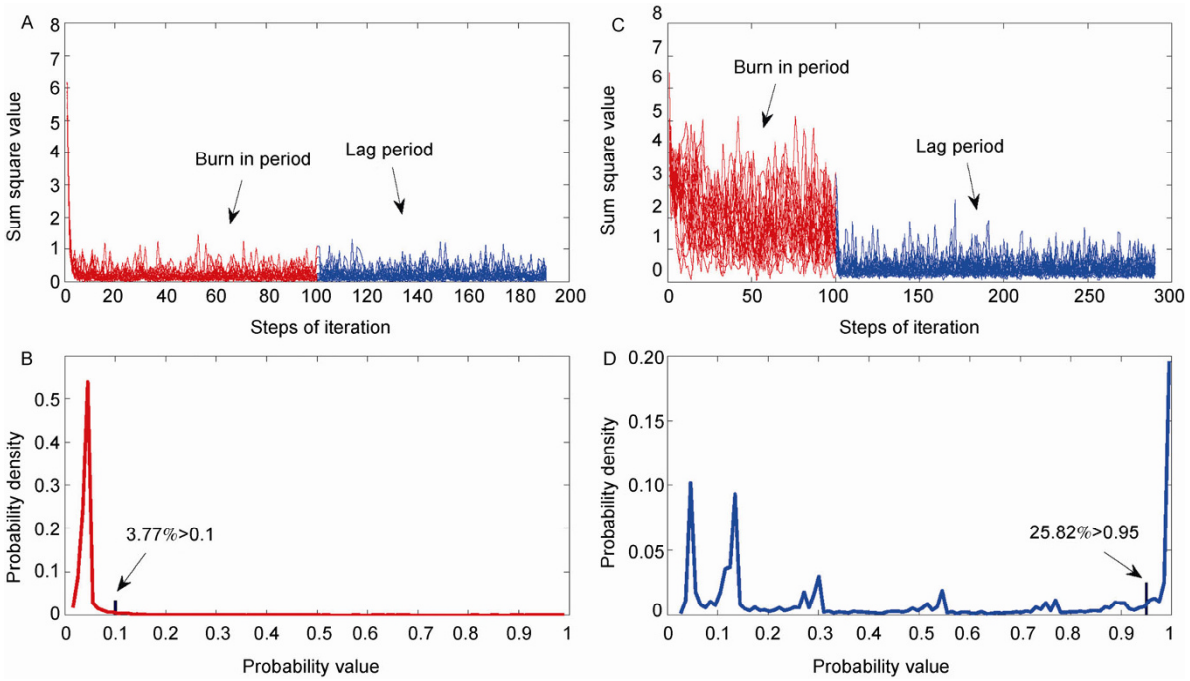

Figure 1 Analyses of stability and reliability of MRF methods (by using single HPRD PPI network for endocrine disease class). A, The variation of posterior probabilities over iteration steps of the kernel-based MRF method. B, The posterior probability distribution of the kernel-based MRF method. There are only 3.77% of unknown vertices which are predicted with probability larger than 0.1, which means only a small amount of significant vertices are predicted with higher probabilities. C, The variation of posterior probabilities over iteration steps of the MRF-Deng method. D, The posterior probability distribution of MRF-Deng method. There are almost 25.82% of unknown vertices that are predicted with probability larger than 0.95, which means too many vertices are predicted with very high probabilities.

2.2 Comparisons between different kernels

To test the contribution of graph kernels in the kernel based MRF method, three types of kernels are employed in our experiments. Figure 2 illustrates the cross-validation results in terms of ROC curves and the AUC score by integrating only three PPI networks and all five biological networks, respectively. The LED kernel achieves the best performance (AUC=0.753) when three PPI networks are integrated, while the MED kernel works best (AUC=0.771) when all five PPI networks are integrated. The similar performance of those kernels also supports the stability of the kernel based MRF method.

Generally, there is no such a kernel that works better than all other kernels in any situation. The LED kernel works better than the MED kernel when three networks are integrated. However, the difference of between those two AUC scores is not large in this situation. Besides, the MED kernel works much better than the LED kernel when five networks are integrated. Hence, the MED kernel is always suggested to be used for multiple data integration if no particular information is obtained.

2.3 Comparisons with existing methods

The kernel-based MRF method is compared with the RWR algorithm, the DIR algorithm, the MRF-Deng algorithm and the IMRF method. Figure 3 illustrates ROC cross-validation results by integrating all five biological networks. It can be seen from the figure that the kernel based MRF method performs best compared with the other four existing algorithms. The kernel-based MRF method achieves the highest AUC score at 0.771 by using the MED kernel, followed by the IMRF method (AUC=0.743), the DIR algorithm (AUC=0.691) and the RWR algorithm (AUC=0.676). The MRF-Deng method achieves the AUC score only at 0.551.

3 Discussion

In this paper, we have presented an improved kernel based MRF method for identifying human disease genes by integrating five biological networks. The presented method is not only flexible in terms of integrating different types of biological data, but also reliable in terms of identifying human disease genes. Three kinds of graph kernels are employed to capture relationships of all vertices based on their global neighborhood characteristics. An improved Gibbs sampling procedure and a novel parameter estimation method are then developed for the presented MRF method. The use of different kernels brings great improvement for the previous MRF method. The proposed MED kernel works similar to the most commonly used LED kernel when three PPI networks are integrated, and it works best when

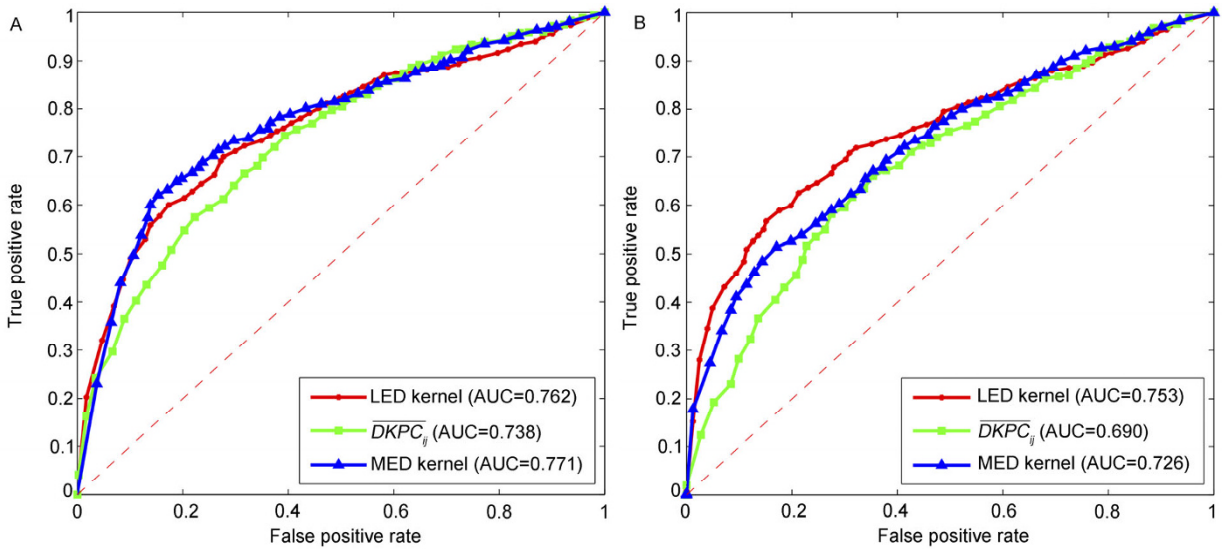

Figure 2 Comparisons of different kernels by using the kernel-based MRF method. A, Comparisons of ROC curves by integrating all five biological networks. B, Comparisons of ROC curves by integrating only three PPI networks. The red lines are ROC curves by using the LED kernels. The green lines are ROC curves by using the $\overline{DKPC}_{ij}$. The blue lines are ROC curves by using the MED kernels. AUC values are listed in parentheses.

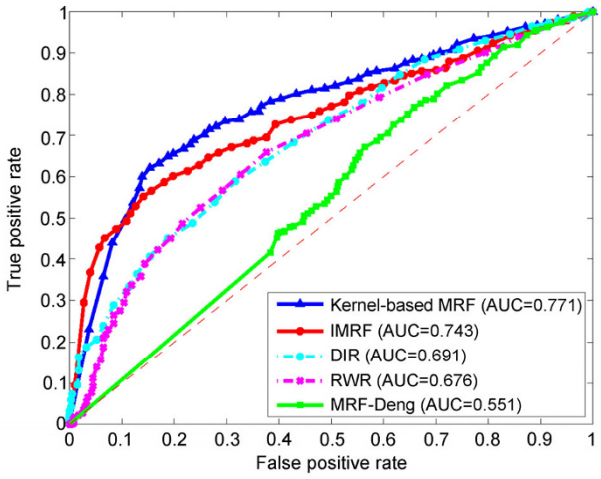

Figure 3 ROC curves of cross-validation results of different methods with integrating five biological networks. The blue solid line represents the ROC curve by using the kernel-based MRF method. The red solid line represents the ROC curve by using the IMRF method. The cyan dash-dot line represents the ROC curve by using the DIR method. The magenta dash-dot line represents the ROC curve by using the RWR method. The green solid line represents the ROC curve by using the MRF-Deng method. AUC values are listed in parentheses.

five biological networks are integrated. Hence, the MED kernel is suggested to be used for the proposed algorithm when multiple data integration is involved to predict disease genes. Predictions by our presented method with integrating all five biological networks achieve the AUC score of 0.771 when the MED kernel is employed, which is very promising for identifying human disease genes.

The authors declare that they have no conflict of interest.

This work was supported by the Natural Sciences and Engineering Research Council of Canada and National Natural Science Foundation of China (61428209, 61232001).

1 Hwang T, Zhang W, Xie M, Liu J, Kuang R. Inferring disease and gene set associations with rank coherence in networks. Bioinformatics, 2011, 27: 2692–2699
2 Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol, 2010, 6: e1000641
3 Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PLoS One, 2009, 4: e4346
4 Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol, 2008, 4: 189
5 Ma X, Lee H, Wang L, Sun F. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. Bioinformatics, 2007, 23: 215–221
6 Lage K, Karlberg EO, Størling ZM, Olason PI, Pedersen AG, Rigina O, Hinsby AM, Tümer Z, Pociot F, Tommerup N, Moreau Y, Brunak S. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol, 2007, 25: 309–316
7 Chen Y, Wang W, Zhou Y, Shields R, Chanda SK, Elston RC, Li J. *In silico* gene prioritization by integrating multiple data sources. PLoS One, 2011, 6: e21137
8 Strohman R. Maneuvering in the complex path from genotype to phenotype. Science, 2002, 296: 701–703
9 Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. J Comput Biol, 2003, 10: 947–960
10 Deng M, Chen T, Sun F. An integrated probabilistic model for functional prediction of proteins. J Comput Biol, 2004, 11: 463–475
11 Kourmpetis YA, van Dijk AD, Bink MC, van Ham RC, ter Braak CJ. Bayesian Markov random field analysis for protein function prediction based on network data. PLoS One, 2010, 5: e9293
12 Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. OMICS, 2006, 10: 40–55
13 Deng M, Tu Z, Sun F, Chen T. Mapping gene ontology to proteins based on protein-protein interaction data. Bioinformatics, 2004, 20: 895–902

14 Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics, 2003, 19: i197–i204
15 Wei Z, Li H. A Markov random field model for network-based analysis of genomic data. Bioinformatics, 2007, 23: 1537–1544
16 Kondor RI, Lafferty J. Diffusion kernels on graphs and other discrete input spaces. In: Proceedings of the Nineteenth International Conference on Machine Learning, San Mateo, USA, 2002. 315–322
17 Ma X, Chen T, Sun F. Integrative approaches for predicting protein function and prioritizing genes for complex phenotypes using protein interaction networks. Brief Bioinform, 2014, 15: 685–698
18 Schölkopf B, Tsuda K, Vert JP. Kernel Methods in Computational Biology. Cambridge: The MIT Press, 2004
19 Chen B, Wang J, Wu FX. Prioritizing human disease genes by multiple data integration. In: IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, 2013. 621
20 Chen B, Wang J, Li M, Wu FX. Identifying disease genes by integrating multiple data sources. BMC Med Genomics, 2014, Suppl2: S2
21 Li SZ. Markov Random Field Modeling in Image Analysis. 3rd ed. Berlin Heidelberg: Springer, 2009
22 Besag J. Spatial interaction and the statistical analysis of lattice systems. J Royal Statist Soc B, 1974, 36: 192–236
23 Kolaczyk ED. Statistical Analysis of Network Data. Berlin Heidelberg: Springer, 2009
24 Kamberova G. Markov random field models: a Bayesian approach to computer vision problems. Department of Computer & Information Science Technical Reports, University of Pennsylvania, 1992
25 Suess EA, Trumbo BE. Introduction to probability simulation and Gibbs sampling with R. New York: Springer, 2010
26 McKsick VA. Mendelian inheritance in man and its online version, OMIM. Am J Hum Genet, 2007, 80: 588–604
27 Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. Proc Natl Acad Sci USA, 2007, 104: 8685–8690
28 Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes HW. CORUM: the comprehensive resource of mammalian protein complexes-2009. Nucleic Acids Res, 2010, 38: D497–D501
29 Kikugawa S, Nishikata K, Murakami K, Sato Y, Suzuki M, Altaf-Ul-Amin M, Kanaya S, Imanishi T. PCDq: human protein complex database with quality index which summarizes different levels of evidences of protein complexes predicted from H-invitational protein-protein interactions integrative dataset. BMC Syst Biol, 2012, 6: S7
30 Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A, Balakrishnan L, Marimuthu A, Banerjee S, Somanathan DS, Sebastian A, Rani S, Ray S, Harrys Kishore CJ, Kanth S, Ahmed M, Kashyap MK, Mohmood R, Ramachandra YL, Krishna V, Rahiman BA, Mohan S, Ranganathan P, Ramabadran S, Chaerkady R, Pandey A. Human protein reference database-2009 update. Nucleic Acids Res, 2009, 37: D767–772
31 Stark C, Breitkreutz BJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res, 2006, 34: D535–539
32 Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct—open source resource for molecular interaction data. Nucleic Acids Res, 2007, 35: D561–565
33 Zhao B, Wang J, Li M, Wu, FX, Pan, Y: Detecting protein complexes based on uncertain graph model. IEEE/ACM Trans Comput Biol Bioinform, 2014, 11: 486–497
34 Wang J, Li M, Chen J, Pan Y. A fast hierarchical clustering algorithm for functional modules discovery in protein interaction networks. IEEE/ACM Trans Comput Biol Bioinform, 2011, 8: 607–620
35 Li M, Wu X, Wang J, Pan Y. Towards the identification of protein complexes and functional modules by integrating PPI network and gene expression data. BMC Bioinformatics, 2012, 13: 109
36 Li M, Chen J, Wang J, Hu B, Chen G: Modifying the DPClus algorithm for identifying protein complexes based on new topological structures. BMC Bioinformatics, 2008, 9: 398
37 Wang J, Li M, Wang H, Pan, Y: Identification of essential proteins based on edge clustering coefficient. IEEE/ACM Trans Comput Biol Bioinform, 2012, 9: 1070–1080
38 Li M, Zheng R, Zhang H, Wang J, Pan Y. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. Methods, 2014, 67: 325–333
39 Tang X, Wang J, Zhong J, Pan Y. Predicting essential proteins based on weighted degree centrality. IEEE/ACM Trans Comput Biol Bioinform, 2014, 11: 407–418
40 Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30
41 Vastrik I, D'Eustachio P, Schmidt E, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. Genome Biol, 2007, 8: R39
42 Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther, 2012, 92: 414-417
43 Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. PID: the pathway interaction database. Nucleic Acids Res, 2009, 37: D674–679
44 Wu C, Orozco C, Boyer J, Leglise M, Goodale J, Batalov S, Hodge CL, Haase J, Janes J, Huss JW 3rd, Su AI. BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. Genome Biol, 2009, 10: R130
45 Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA, 2004, 101: 6062–6067
46 Köhler S, Bauer S, Horn D, Robinson PN. Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet, 2008, 82: 949–958